Comparative Study of BERT-CNN, TRANS-BLSTM, and RoBERTa Models for Sentiment Analysis

1st Christopher Alden Anugrah Silitonga
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
christopher.silitonga@binus.ac.id

2nd Marco Davincent Dermawan
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
marco.dermawan@binus.ac.id

3rd Franz Adeta Junior
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
franz.junior@binus.ac.id

4th Nadia Nadia
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nadia002@binus.ac.id

***Abstract*—In the field of Natural Language Processing (NLP), various milestones and improvements have been made due to the advent of deep learning, mainly in sentiment analysis of text data. BERT has excelled through its capability to comprehend emotional context. Nonetheless, there are some limitations with BERT like long inference times as well as interpretability challenges. Therefore, combinations of such architectures with BERT, for instance TRANS-BLSTM, BERT-CNN and RoBERTa have been studied. These hybrid models aim at leveraging on the strengths that BERT possesses while making sentiment analysis tasks more efficient and interpretable. This investigation demonstrates ongoing attempts to develop the better NLP models that can comprehend human emotions reflected in text data more effectively.**

Keywords*— *deep learning, emotion recognition, natural language processing, sentiment analysis transformers

I. Introduction

The rise of digital platforms has generated a vast amount of text data from social media, forums, news, and other sources. Extracting emotional context from this data is challenging but crucial for understanding the sentiments that drive decision-making, especially in business [1]. Sentiment analysis is essential for measuring customer satisfaction, tailoring marketing campaigns, and making informed recommendations based on consumer opinions. Customer perceptions can generally be classified as positive, negative, or neutral.

Earlier Natural Language Processing (NLP) models, such as Naïve Bayes and Support Vector Machines (SVM), were primarily probabilistic. These models struggled with capturing context, ambiguity, and the intricacies of human language, leading to unsatisfactory sentiment analysis outputs [2]. For instance, they had difficulty handling irony, sarcasm, and complex word relationships.

NLP has been transformed by transformers like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and RoBERTa (Robustly optimized BERT approach) with significantly better contextual understanding. Transformers use self-attention mechanism to capture word relationships within a text without relying on sequence order, thus they are more efficient in dealing with long-term dependencies. For instance, OpenAI's GPT model is built upon the transformer architecture to generate coherent and contextually relevant text [3], therefore highlighting how this model understands and produces human-like language. On the other hand, practical issues like high computational costs and slow fine-tuning process make it difficult to employ BERT for real-world applications.

This study compares advanced models like BERT-CNN, RoBERTa, and Trans-BLSTM which were were chosen for their improved contextual representation on identical datasets to determine a model best-balancing accuracy against computational efficiency. It aims at giving an overview about the weaknesses and strengths of individual models thus making them more effective in NLP development as well as better business strategies.

II. Related Works

A. Sentiment Analysis

Sentiment Analysis [4] through Natural Language Processing (NLP) has been refined over the years. By its core NLP combines computational linguistics, rule-based language models, and statistical methods with machine learning, such that computers have the ability to comprehend, interpret, and generate human language in a manner that is meaningful while also valuable. Detection of emotions from a text or long-length sentence through sentiment analysis and emotion detection represents pivotal advancement. Each textual piece harbors multitude of emotions, and discerning these emotions hold immense potential in a lot of diverse sectors. Leveraging technique like deep learning, can be refined through the extraction of article, news, social media that can be training datasets to chance the accuracy and bolstering its classification capabilities [5]

B. LSTM Networks: Pioneering Long-Term Dependency Modeling in NLP

Traditional RNNs struggle with long-term dependencies due to vanishing gradient problem during training, which LSTM is designed to overcome. This is possible from the use of gated cell structures which regulates the flow of information allowing the network to either remember or forget. This have proven to be highly effective for tasks that require understanding of sequential data, such as; sentences and speech, as in language processing [6].

Building upon the established LSTM network, Bidirectional LSTM (Bi-LSTM) neural network incorporate a forward pass and backward pass through the input data. This enables model to have insights from both past and future at any point of the sequence, which contributes to helping the model's ability to understand context.

In practical applications, such as the analysis conducted on the Amazon Product Review dataset [7] and opeNER Datasets Bi-LSTMs have demonstrated superior capability in classifying sentiments as positive or negative by BILSTM 83 and LSTM 83.1 [8].

C. Evolution of Transformer-Based Models in NLP: Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) has become a transformative force in Natural Language Processing (NLP) [9]. Developed by Google Research, BERT's bidirectional training of Transformers allows it to understand both the left and right context of a token within a sentence, enhancing its ability to handle complex NLP tasks like question answering, language inference, and sentiment analysis [9].

BERT's innovation lies in its pre-training strategies: masked language model (MLM) and next sentence prediction (NSP) [9]. MLM masks words in the input and predicts them based on context, while NSP predicts whether two text segments are sequential, improving its understanding of text relationships.

BERT's versatility is clear in its application across various NLP tasks with minimal task-specific modifications [10]. It adapts seamlessly to sentence-level tasks like classification and entailment, as well as token-level tasks like named entity recognition.

BERT has also been foundational in developing specialized models like BERT-LSTM, BERT-CNN, and hybrids of Bi-LSTM and CNN, which leverage its deep contextual embeddings for even more complex tasks, as discussed in this paper [11].

D. Advancing BERT's Architecture: RoBERTa

The robust optimization approaches of RoBERTa, also known as the robustly Optimized BERT Approach proposed by Liu et al. [12], have drawn attention as an extension of BERT. It modifies several key hyperparameters and training strategies to enhance performance across multiple NLP tasks. This model was trained without the next sentence prediction objective, and instead, it was trained on longer sequences, larger batches, and more extensive datasets which totalled over 160GB of text, including Stories, CC-News, BooksCorpus, and OpenWebTe [12]. It was trained

Performance wise, RoBERTA achieved a new SOTA result across multiple benchmarks. On the GLUE benchmark, it received 88.5%, surpassing BERT's previous best. Its overall accuracy rose to 83.2% outperforming BERT's 72%. And additionally on the SQuAD v2.0 roBERTa reached an F1-score of 89.4% [12].

E. Architecture Integrations

Recent advances in NLP have been driven by integrating deep learning architectures. The TRANS-BLSTM model, developed by Huang et al. (2020) [13], combines Transformer and BLSTM, excelling in sequential data processing and capturing contextual information from both past and future inputs.

This model addresses the limitations of BERT's reliance on MLM, which can reduce its effectiveness across varied NLP tasks. On the SQuAD 1.1 development dataset, TRANS-BLSTM achieved an F1 score of 94.01%, outperforming models like BERT and TRANS/BERT [13].

Another innovation is BERT-CNN, which integrates transformer-based models with CNNs to enhance emotion detection and classification from text, as introduced by Abas et al [14]. BERT captures contextual dependencies, while the CNN enhances feature extraction.

Empirical results show BERT-CNN outperforms traditional methods, achieving 94.7% accuracy and a 94% F1-score on the semeval2019 task3 dataset, and 75.8% accuracy with a 76% F1-score on the ISEAR dataset [14].

F. Challenges in Advanced NLP Model Integration

Several challenges highlighted in foundational literature, such as *Natural language processing: state of the art, current trends and challenges* [15], remain. These challenges are crucial for further refinement and broader application of these technologies.

1) Model Generalization: Despite achieving high accuracy on benchmark datasets, there is a noted challenge in generalizing these results to real-world datasets or across diverse domains [15]. Models trained on specific types of data often fail to perform well on others that differ in style, context, or inherent characteristics.

2) Complexity in Integration: The integration of different complex models such as LSTMs with Transformers introduces additional complexity in terms of model architecture and parameter tuning. This complexity can make these models less accessible to practitioners without deep expertise in machine learning [15].

3) Bias and Fairness: There is also an ongoing concern with bias in machine learning models. Advanced NLP models, trained on large but often biased data sets, can perpetuate or even amplify these biases in their outputs, leading to fairness issues in their applications [15].

4) Robustness and Security: The robustness of these models against adversarial attacks and other security vulnerabilities is another area of concern. Ensuring that these models can withstand malicious inputs and provide reliable outputs is essential for their safe deployment in critical systems [15].

III. Methodology

This section will further explore assessing new versions of BERT models on a sentiment analysis dataset. In particular, it will focus on training TRANS-BLSTM, RoBERTa, and BERT-CNN models, obtaining their outputs, and analyzing them for insights.

A. Data Description

The dataset utilized in this study was taken from the Amazon Fine Food Reviews dataset which consists of user-generated reviews of various food products available on Amazon. The time range of the reviews is from October 1999 to October 2012, totaling 568,454 reviews, there are 393,579 unique reviews, 256,059 users, and 74,258 products. Notably, 260 users have written more than 50 reviews each. The dataset includes various features such as product details, user information, profile names, helpfulness ratings, scores, timestamps, summaries, and the main review text.

B. Data Cleaning and Preprocessing

A subset of 30000 rows was used in the process, each entries was ensured to be exactly the same for each iteration minimizing the disturbance of the end result. The data was equally divided for each class which are positive, neutral, and negative, with a fixed random state to ensure consistency across experiments.

The data was then processed to an input the model could understand. This step involved mapping numerical ratings against every review to sentiment labels. In particular, ratings less than 3 were considered negative sentiment and represented by 0, ratings equal to 3 were represented by 1, regarded as neutral, and ratings above 3 were represented by 2 as positive.The processed dataset was split into three sets: a training set, a test set, and a validation set, with a division ratio of 70:20:10, respectively. I.

TABLE I: Distribution of classes, number of reviews, vocabulary size, median review length of the datasets

Dataset	Classes	Number of reviews	Vocabulary size	Median of review length
Original	3	568454	586944	302
Training	3	21000	84138	344
Testing	3	6000	41262	342
Validation	3	3000	27345	347

C. Model Architecture

RoBERTa, BERT-CNN, and TRANS-BLSTM have been chosen for their differences in the architectural approaches, providing a comprehensive comparison and insight of how different alterations will affect the model. RoBERTa will serve as a baseline model for benchmarking other models.

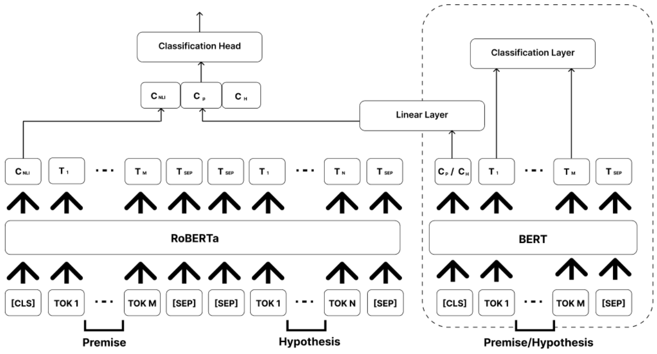

Fig. 1: RoBERTa Architecture

1) RoBERTa: operates using multiple layers of Transformer encoders. It utilizes a self-attention mechanism, which allows the model to weigh the importance of different words in a sequence shown in Fig. 1, enhancing contextual understanding. Unlike its predecessor BERT, RoBERTa focuses solely on the Masked Language Modeling (MLM) objective, removing the Next Sentence Prediction (NSP) task. It is trained on a significantly larger dataset and for longer periods, with adjustments to key hyperparameters to improve performance.

Given an input sequence

$$X = [x_1, x_2, \ldots, x_n] \tag{1}$$

self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2}$$

Here, Q, K, and V are the query, key, and value matrices derived (2) from the input X (1), respectively, and d_k denotes the dimension of the key and query vectors.

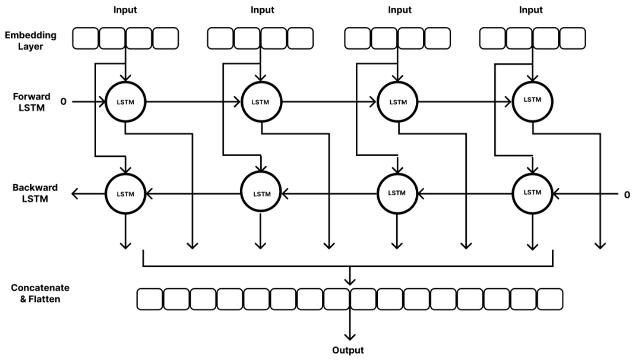

Fig. 2: BLSTM Architecture

2) BLSTM: works by processing input sequences through two LSTM layers shown in Fig. 2: one moving forward and the other moving backward. This bidirectional processing allows the model to capture context from both past and future tokens. The architecture combines multiple LSTM cells, each maintaining its own cell state and hidden state, which helps in preserving and transferring information over long sequences. This approach makes BLSTM particularly effective for tasks involving sequential data, such as time series analysis or natural language processing.

LSTM Cell: Given input x_t, previous hidden state h_{t-1}, and previous cell state c_{t-1}:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{c}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

where σ is the sigmoid function and $\odot$ denotes element-wise multiplication (3).

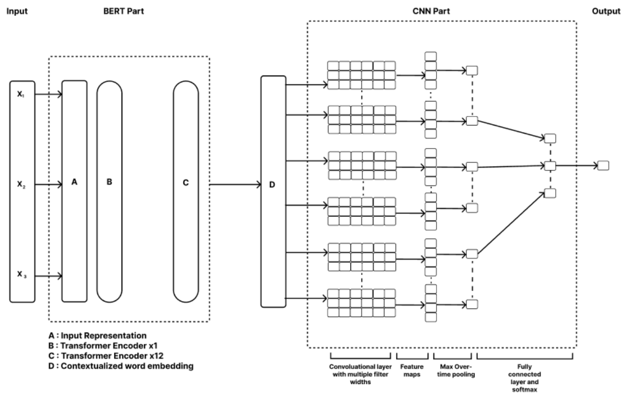
Input
BERT Part
CNN Part
Output
X₁
X₂
X₃
A
B
C
D
A : Input Representation
B : Transformer Encoder x1
C : Transformer Encoder x12
D : Contextualized word embedding
Convoluational layer with multiple filter widths
Feature maps
Max Over-time pooling
Fully connected layer and softmax

Fig. 3: BERT-CNN Architecture

3) BERT-CNN: Operates by leveraging BERT to generate contextualized word embeddings for the input text, capturing subtle meanings. There are also convolutional layers to extract local features from the text which are followed by a pooling layer that reduces the dimensionality of the feature maps while keeping track of the important features. Such a combination inherits the contextual understanding of BERT and the feature extraction capabilities of CNN. Given an input matrix X and a filter W:

$$Z_{i,j} = \sum_{m=1}^{k} \sum_{n=1}^{k} X_{i+m-1,j+n-1} \cdot W_{m,n} \tag{4}$$

where k is the filter size and Z is the output feature map (4).

The BERT-CNN was initialized with the pretrained "bert-base-uncased" weights. The BERT model has a hidden size of 768, 12 transformer blocks, 12 self-attention heads, and a vocabulary of 30,000 trained on the BookCorpus. A 1-dimensional convolutional layer was added with an input channel size of 768 and an output channel size of 128. The kernel size was set to 5 with padding of 2. To reduce dimensionality, a Max-pooling layer was added, followed by a dropout layer with a rate of 0.5 to prevent overfitting. The high-level features were then fed into a fully connected layer with a softmax activation function. The Adam optimizer and Categorical cross-entropy loss function were used. The architecture is shown in Fig. 3, and the parameters are listed in Table II.

TABLE II: Hyper-parameters used for each model

Hyper-parameters	Values
Learning rate	4e-5
Loss function	Categorical Cross-entropy
Optimizer	Adam
Dropout	0.5
Kernel sizes	5
Epochs	3

D. Training

The training was done on Kaggle Notebook with the following specifications shown in Table III

TABLE III: Hardware and software specifications used for training

Hardware	Specifications
CPU	Intel(R) Xeon(R) CPU @ 2.00GHz
RAM	31.36 GB
GPU	NVIDIA Tesla P100 16 GB
OS	Ubuntu 22.04.3

Initially, the data undergoes a tokenization process, converting it into a format recognizable by the BERT model as input. Each model was trained with the same training arguments, as shown in Table IV.

TABLE IV: Training arguments used for each model

Training arguments	Values
Epochs	3
Batch size for training	16
Batch size for evaluation	64
Warmup steps	500 steps
Weight decay	0.01
Evaluation strategy	steps
Frequency of evaluation	500 steps
Load best model	True
Metrics for best model	f1

Each model was trained using associated specified hyperparameters when discussing the models architecture above. The idea is to maximize the potential of each model while training it in the same environment. After training, the models were tested using a separate script that evaluates their performance on the test set by comparing the prediction results to the ground truth to obtain metrics.

E. Evaluation

A few metrics are taken account for consideration while using the F1-score as a primary key metric.

1) Precision (5): Calculates the ratio of correctly predicted positive instances to the total number of predicted positive instances. Its useful in measuring the models capability in identifying true positives but is sensitive to class imbalances favoring the majority.

2) Recall (6): Calculates the ratio of correctly predicted positive instances to the total number of actual positive instances. Its useful in capturing the models ability in identifying relevant true positives but is sensitive to class imbalances.

3) F1-score (7): Balances Precision and Recall by providing a single metric that considers both.

4) Accuracy (8): Calculates the ratio of correctly predicted instances to the total number of instances. While easy to interpret it does not consider the distribution of errors.

The formula used to calculate the respective scores are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Finally, we calculate each of the metrics for each of the models mentioned here (RoBERTa, TRANS-BLSTM, BERT-CNN), and from the scores compare which of the model as best for sentiment classification for our particular dataset.

IV. Result

This section will primarily discuss the training process outcome, results, and output of each model.

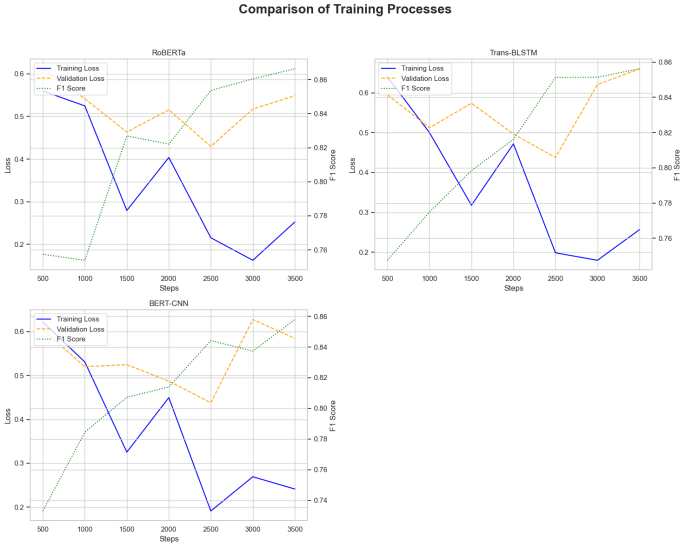

Fig. 4: RoBERTa, BERT-CNN, and TRANS-BLSTM Training Validation Loss over Steps with F1-Score

A. Training process

From Fig. 4 Observing the training loss, after the initial decrease, fluctuations and occasional increases can be seen. Although this might suggest that the model is improving on the training set, it is not necessarily a positive indicator overall. The validation loss remains relatively stable with a slight dip in the middle. For example, in the case of RoBERTa, the validation loss starts at 0.611742 and settles at 0.548481 towards the end. This outcome is neutral, but ideally, a consistent decrease in validation loss would indicate better performance on unseen data indicating that it is learning generalizable patterns rather than simply memorizing the training dataset, which is the desired outcome. The F1-Score for each model shows a significant increase, which generally suggests that the model is improving in its performance.

Overall, the graph suggest that the training process could benefit from better regularization techniques or a more optimized learning rate to enhance the model's generalization capabilities.

B. Model Output

TABLE V: Confusion Matrix Results for RoBERTa, BERT-CNN,and TRANS-BLSTM Models

Model	Sentiment	TP	TN	FN	FP
RoBERTa	Negative	1659	3710	347	284
	Neutral	1714	3472	320	494
	Positive	1738	3929	222	111
BERT-CNN	Negative	1595	3761	411	233
	Neutral	1741	3419	293	547
	Positive	1724	3880	236	160
TRANS-BLSTM	Negative	1712	3662	294	332
	Neutral	1628	3606	406	360
	Positive	1780	3852	180	188

Table V shows the output given by respective models. A common motif shown from the data is:

1) Overall, each class within sentiment category is relatively balanced with other classes.
2) A larger amount of false prediction within the neutral sentiment category compared to the total instances in this class.

C. Comparative Analysis on Model Performance

TABLE VI: Comparative Analysis Results for RoBERTa, BERT-CNN, and TRANS-BLSTM Models

Model	**Sentiment**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
RoBERTa	Negative	0.8538	0.8270	0.8402	0.8948
	Neutral	0.7763	0.8427	0.8081	0.8643
	Positive	0.9400	0.8867	0.9126	0.9445
	Average			**0.8536**	
BERT-CNN	Negative	0.8725	0.7951	0.832	0.8927
	Neutral	0.7609	0.8559	0.8056	0.86
	Positive	0.9151	0.8796	0.897	0.934
	Average			**0.8448**	
TRANS-BLSTM	Negative	0.8376	0.8534	0.8454	0.8957
	Neutral	0.8189	0.8004	0.8095	0.8723
	Positive	0.9045	0.9082	0.9063	0.9387
	Average			**0.8537**	

Evaluation metrics for each model can be seen in Table VI which are computed from the models output in Table V.It's worth acknowledging that these models are only trained with 21,000 rows of data and given 3 epochs due to computational resource constraints. This might reflect on the results where the differences across model outputs might not be prominent.

1) RoBERTa: Generally outperforms BERT-CNN in every class, particularly with better recall and precision in the neutral class. The performance boost comes from the transformer architecture, which helps in understanding nuanced sentiments. RoBERTa's extensive pretraining on the BookCorpus with a vocabulary of 30,000 aids its ability to generalize and understand broader language patterns.

2) BERT-CNN: Shows the lowest performance among the three models. Since this architecture uses BERT to produce features, which are then classified by the CNN layer, it might be less effective in capturing long-range dependencies than conventional transformer architectures.

This inability to capture long-range dependencies results in higher false negatives and false positives in the neutral class. Considering that the median review length is around 300 words, it can miss context and nuances, as seen in the neutral class output, which is the worst among the three.

3) TRANS-BLSTM: Performs the best among the three models, particularly in the positive and neutral classes, showing the highest true positive rates and lower false negatives. It balances precision and recall effectively across all classes.

The transformer component captures the global context, while the LSTM component is effective at sequential dependencies. Combining both leverages the strengths of each architecture, helping maintain a comprehensive understanding of the text, especially in handling subtleties in neutral sentiment, outperforming both BERT-CNN and RoBERTa.

V. Conclusion and Future Works

For applications like business decisions and emotion analysis, choosing the correct model architecture is very important. TRANS-BLSTM performs best in terms of positive and neutral categories with higher true positive identification than others do. It combines strengths of transformers and LSTMs to effectively balance between precision and recall specifically for better discrimination of neutral sentiment compared to BERT-CNN and RoBERTa.

However, RoBERTa always outperforms BERT-CNN in neutral, having superior recall and precision associated with its optimal transformer design as well as extensive pretraining it has undergone. The hybrid structure of BERT-CNN, on the other hand, makes it difficult for it to adequately capture long-range dependencies thus leading to more misclassification errors especially at the level of neutral category.

Future research should focus on improving model efficiency, refining validation, and developing ethical NLP models. This will help maximize the potential of current and future NLP technologies by encouraging innovative solutions to practical problems such as sentiment analysis or otherwise.

References

[1] N. D. Tselikas and D. K. Nasiopoulos, "Llms and nlp models in cryptocurrency sentiment analysis: A comparative classification study," 2024.

[2] D. M. Anindya Nag and G. Mobin, "A comparison of the several speech tagging models used in nlp," 2024.

[3] S. K. T. J. d. S. J. Dr. Pankaj Malik, Palak Khatri, "Analyzing the robustness of nlp models to diverse prompts," 2024.

[4] L. T. C. Ottoni, A. L. C. Ottoni, and J. de Jesus Fiais Cerqueira, "A deep learning approach for speech emotion recognition optimization using meta-learning," *Electronics (Switzerland)*, vol. 12, 12 2023.

[5] B. Liu, "Asentiment analysis and subjectivity," 2010.

[6] H. E.-A. A. K. Ngoc Tran Khanh Le, Nadia Hadiprodjo and A. Teshebaev, "The recent large language models in nlp," 2023.

[7] U. B. Mahadevaswamy and P. Swathi, "Sentiment analysis using bidirectional lstm network," vol. 218. Elsevier B.V., 2022, pp. 45–56.

[8] R. K. BJeremy Barnes and S. S. im Walde, "Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets," 2017.

[9] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "Bert: Pre-training of deep bidirectional transformers for language understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[10] K. van Deemter, "Dimensions of explanatory value in nlp models," 2022.

[11] G. H. S.-u. K. Beakcheol Jang, Myeonghwi Kim and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," 2020.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 7 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[13] Z. Huang, P. Xu, D. Liang, A. Mishra, and B. Xiang, "Trans-blstm: Transformer with bidirectional lstm for language understanding," 3 2020. [Online]. Available: http://arxiv.org/abs/2003.07000

[14] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "Bert-cnn: A deep learning model for detecting emotions from text," *Computers, Materials and Continua*, vol. 71, pp. 2943–2961, 2022.

[15] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, pp. 3713–3744, 1 2023.